DoD Executive Agent
Office of the
Assistant Secretary
of the Army
(Installations and
Environment)

# Advanced Internet Data Search Portal for Environmental Applications

## Joint Services Environmental Management Conference
## May 21-24, 2007

Andy Del Collo
CNO Environmental Readiness Division (N45)

*The views, opinions, and/or findings contained in this briefing are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other official documentation. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products. All product names and trademarks cited are the property of their respective owners.*

The NDCEE is operated by: CTC Concurrent Technologies Corporation

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

| | | |
|---|---|---|
| 1. REPORT DATE<br>**MAY 2007** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-00-2007 to 00-00-2007** |

| | |
|---|---|
| 4. TITLE AND SUBTITLE<br>**Advanced Internet Data Search Portal for Environmental Applications** | 5a. CONTRACT NUMBER |
| | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**Chief of Naval Operations (CNO) ,Environmental Readiness Division (N45),2000 Navy Pentagon,Washington,DC,20350-2000** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR’S ACRONYM(S) |
| | 11. SPONSOR/MONITOR’S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

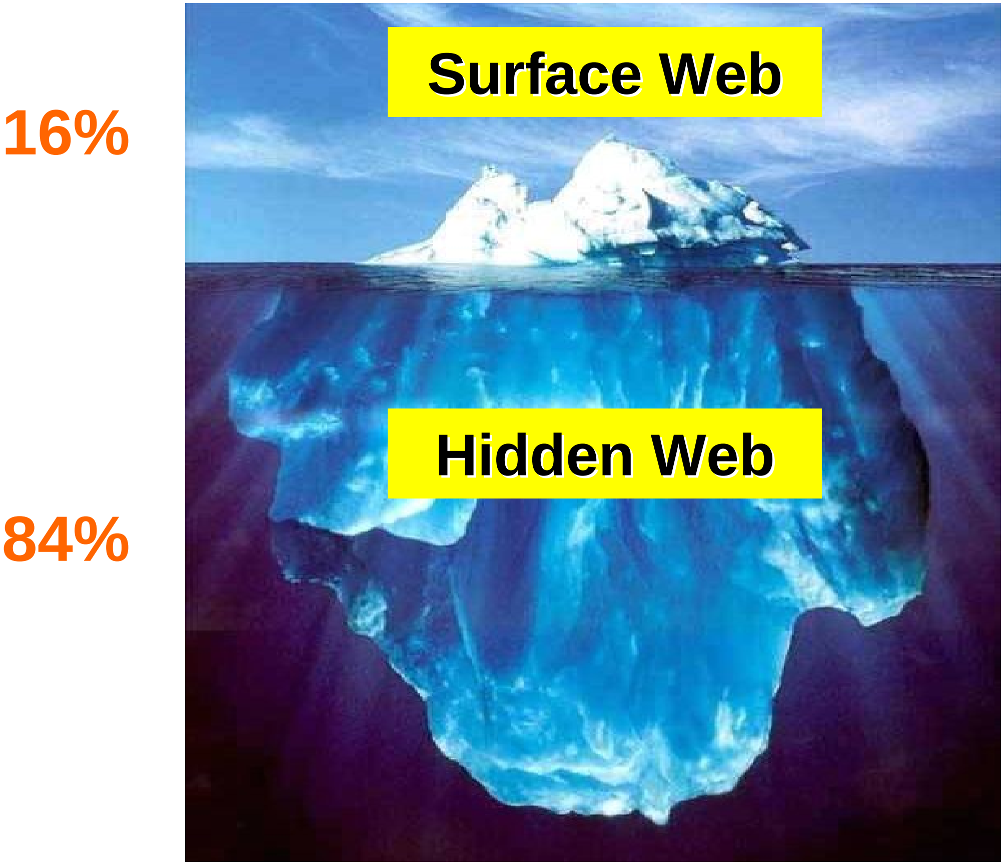
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **15** | |

# Motivation for Project

- Effective Government decision-making for Environmental, Safety and Occupational Health (ESOH) technology investments and operations requires current information on many issues, including:
  - Related technology developments and trends
  - Active organizations / individuals
- Massive amounts of relevant information exist on the web, but are scattered across many sources and are often not readily available
  - Technology press releases
  - Technical and trade journals
  - Organization web sites
  - Conference proceedings
- Gathering and understanding this data is a significant challenge for program managers, technology developers, technology purchasers, and other decision-makers

# The Problem

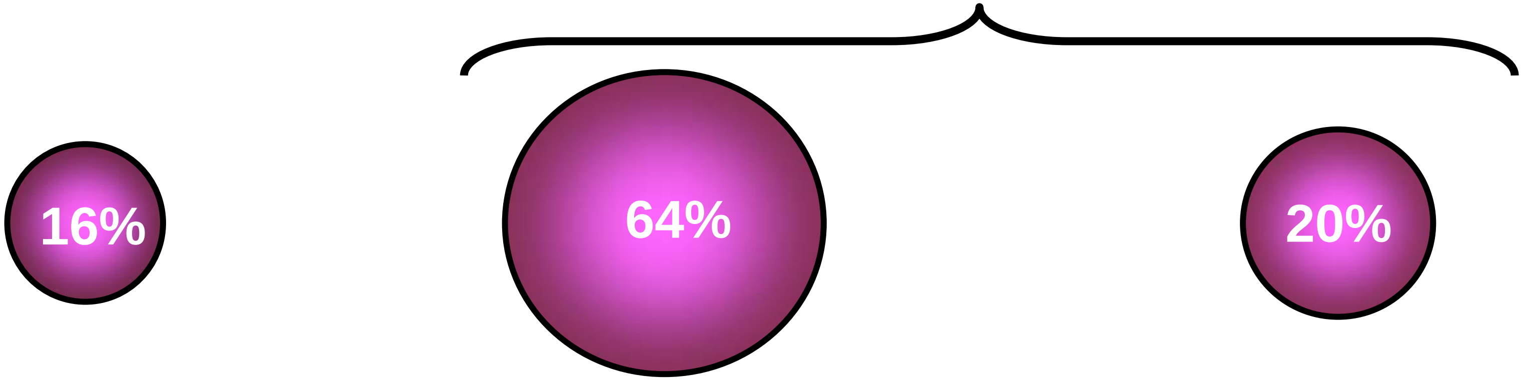
General search engines index only the “Surface Web” – only reaching 16% of the available information



# Categories of Web Information

## SURFACE WEB

## HIDDEN WEB



16%

64%

20%

- **Free Access**
  - Indexed by search engines such as Google

- **Free Access**
  - **Not** indexed by traditional search engines
  - Need to know where to look and how to search

- **Restricted Access**
  - Paid Subscription or Registration
  - **Not** indexed by traditional search engines
  - Need to know where to look and how to search

# Deep Web Examples

| Resource Category | Sample Sites | Find Using Google? | Find with Deep Web Search? |
|---|---|---|---|
| **New Product Announcements** – Sites that publish press releases, product announcements or featured products | GlobalSpec – Product Announcements<br>Reed Business Information | No | Yes |
| **Trade Journals** – Journals devoted to specific applications, typically include product directories, product announcements, articles by company representatives | Pennwell – Water & Wastewater International<br>Paint & Coatings Industry Magazine | No | Yes |
| **Technical Literature** – Technical journals containing articles from universities and companies | SciRus<br>SciTation<br>Google Scholar | No | Yes |
| **Product directories** | Thomas Register<br>Global Spec – product directory | Limited | Yes |
| **Funded Research** | EPA Funded Research Clu-In | No | Yes |

# The Benefits of a Deep Web Search Portal

- Opening up relevant portions of the hidden Web (84% of the total web) to routine searching – would allow a consistent set of resources to be available to all users
- More targeted searching of the best sites – site source list updated by site manager so new sites become available to all users as they are identified
- Automatic quick searches of the best sites for priority topics and materials needed – user does not need to learn the search commands for several different sites
- Faster searches of relevant Web sites for environmental technologies – cuts wasted time finding sites
- On-line ability to organize and screen post-search results – cuts wasted time to evaluate initial search results
- Reduced risk of missing key information that could affect future environmental actions

# Task Objectives

1. Develop and test an advanced, strategic data search and analysis system prototype based on commercial-off-the-shelf (COTS) tools for online searching

2. Identify a Department of Defense (DoD)-wide group of potential system users and other stakeholders to test the prototype design and support operational implementation

3. Develop the business model for operational system deployment and the roadmap to full implementation for DoD users

# Objective 1 Accomplishments- Prototype System

- **Completed User Needs Assessment**
- **Completed Prototype System Design**
  - Screened 34 COTS deep web search engines
  - Evaluated 7 leading COTS deep web search engines
  - Selected 2 COTS vendors for demonstration
    - Bright Planet (***http://www.brightplanet.com/***)
    - Deep Web Technologies (***http://www.deepwebtech.com/)***
- **Developed Prototype Advanced Web Search Portal**
- **Completed Proof of Concept Demonstration/ Validation for the two candidate systems**
- **Identified a third candidate system with potentially superior capabilities at reasonable cost –** Exalead ***(http://www.exalead.com/search)***

# Deep Web Mining Approaches

| APPROACHES | HARVESTER | FEDERATED | CRAWLER |
|---|---|---|---|
| **DATA COLLECTION METHOD** | Collects materials from identified sites based on defined filter, reindexes materials for additional search capabilities | Connections are configured for defined Web sites, search is run simultaneously on all sites | Indexes designated sites, index identifies key concepts and incorporates additional capabilities |
| **ADVANTAGES** | ■Focused collection reduces search time<br>■Material can be organized into categories for browsing | ■Searches are not restricted by filter concepts<br>■Limited field searching may be available<br>■Results are always current | ■Searches are not restricted by filter concepts<br>■Taxonomy or search suggestions available<br>■Provides both speed and flexibility of search topic |
| **DIS-ADVANTAGES** | Certain sites, such as *GlobalSpec*, cannot be harvested | Quality of search results depend on quality of each site's search engine | *Proof-of-concept not completed* |
| **EXAMPLE SYSTEMS** | BrightPlanet | Deep Web Technologies | Exalead |

# Key Features of the Prototyped Systems

- **Design Concept**:
  - “One-stop shopping” Web portal for searching pre-selected online sites, including deep Web applications, to access ESOH technology, news and other information
  - Efficient, useful responses to straightforward search queries
- **Key Components:**
  - COTS deep Web search tools (three tools selected for evaluation)
  - Search results provided as highly ranked URL locations
  - Search results files can be manipulated for screening and organization of key content
- **Access**: Password-protected access for authorized DoD users
- **User Interface**: User-friendly to accommodate novice searchers

# Objective 2 Accomplishments – Project Stakeholders

| Agency | Organizations | Individuals |
|---|---|---|
| OSD | 5 | 10 |
| Army | 7 | 15 |
| Navy, Marines | 8 | 19 |
| Air Force | 5 | 11 |
| NASA | 1 | 2 |

# Objective 3 Accomplishments- Business Model (Still Being Refined)

- Use DENIX as Search Engine Host, User Access, & Help Desk
- License Deep Web Search Engine via DENIX Program Office
- Use ESOH Advisory Groups to define lists of primary web sites for searches to keep system responsive to their search requirements:
  - PAO Review
  - RDT&E Investment
  - Emerging Contaminants Tracking
  - Equipment/Services Procurement
- Contract for Paid Subscription Sites via DENIX Program Office
- Advisory Group & Search Engine Support provided by NDCEE
- Cost Sharing Between DENIX Program Office & Primary Users
  - Target Using Organization Cost About $50K

# Next Steps

- Plan and execute final COTS system proof-of-concept demonstration (June 2007)
- Final Design Report (July 2007)
- Technology Transfer Information Report: Business model and roadmap to operational system design and use (August 2007)
- Final Task Report (September 2007)

# Acknowledgements

| | |
|---|---|
| **NDCEE Executive Agent** | Mr. Tad Davis, DASA (ESOH) |
| **NDCEE Program Director** | Mr. Hew Wolfe, ODASA (ESOH) |
| **NDCEE Program Manager** | Dr. Charles Lechner, ODASA (ESOH) |
| **NDCEE Contracting Officer's Representative** | Mr. Tom Moran, ODASA (ESOH) |
| **Government Technical Monitor** | Mr. Kurt Buehler, NFESC |
| **NDCEE Project Manager** | Dr. Brad Ashton, *CTC* |

*This work was funded through the Office of the Assistant Secretary of the Army (Installations and Environment) and conducted under contract W74V8H-04-D-0005 Task N.0429-N2.*

# Contact Information

Kurt Buehler
Naval Facilities Engineering Service Center (NFESC)
TEL:  (805) 982-4897
EMAIL:  kurt.buehler@navy.mil

Andy Del Collo
CNO (N45)
TEL:  (703) 602-4497
EMAIL: andy.delcollo@navy.mil

Brad Ashton
Concurrent Technologies Corporation (*CTC*)
TEL:  (703) 310-5653
EMAIL:  ashtonw@ctc.com